VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Data Science

Sanjog Prakash Pawar[1], Manas Nitin Sawant[2]
[1](Master of Computer Application, VIVA institute of Technology/Mumbai University, India)
[2](Master of Computer Application, VIVA institute of Technology/Mumbai University, India)

Abstract : *Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. The term data science was initially used as a substitute for computer science by Peter Naur. In industry data science is often considered as the New Kid on the Block even though some of the data-intensive science such as bioinformatics, the high-energy Physics have been using some sort of data science more than a decade. The Committee on Data for Science and Technology defined data science as the methods and technologies used to conduct scientific research through management and utilization of scientific data. There are many research area such as medical, astrophysics, etc totally based on data science. Data science is related to data mining, machine learning and big data.*

Keywords - *Analytics, Challenges, Data Mining, Data Science, Machine Learning*

I. INTRODUCTION

Data Science is a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract information and data from sound, structured and unstructured data, and utilizes information and data that is possible from data across a wide range of application domains. Data science is related to data mining, machine learning and big data. Data science "is the concept of combining statistics, data analysis, informatics, and their related methods" to "understand and analyze real events" with data. It uses techniques and theories derived from many fields within the context of mathematics, mathematics, computer science, information science, and background information. However, data science is different from computer science and information science.

Data science, a newly developed space with many disciplines related to data collection, processing, analysis, visualization, management and storage of large amounts of information aimed at generating data value itself. The term data science (originally used in exchange for data) was originally used to replace computer sceince by Peter Naur in 1960. In the science data industry it is often referred to as the New Kid on the Block although another data science such as bioinformatics, high-energy Physics has been using some form of data science for more than a decade. The Data Science and Technology Committee (CODATA) has defined data science as the methods and technologies used to conduct scientific research on the management and use of scientific data. Scientific data is highly structured, it is easy to extract information, making analysis easier, more accurate or more accurate. There are many areas of research such as medicine, astrophysics, etc. The latest technology in information technology allows end users to generate big data in companies like Amazon, eBAY, Google or Facebook. This data can be used to predict a new business strategy. For example, Amazon uses integrated filters to produce high quality products to compliment an online customer, Facebook uses the People feature you may know to recommend contacting friends.

1.1 Challenges:

1.1.1 Multiple Data Sources: Companies have begun using various software and mobile applications such as ERP and CRM to collect and manage information related to their customers, sales or employees. Combining data into different, informal or small-scale information can be a complex process. This leads to non-uniform formats as each tool collects information in its own ways. In addition, this also means that there are different sources for managing and extracting data from it.

1.1.2 Data Security: Business data in business is used to identify business opportunities, improve overall business performance and advance decision-making. However, data security is still one of the key issues in data science affecting businesses worldwide. Data security is an umbrella term that encompasses all security measures and tools used in data statistics and processes. A few of the data security violations included

- Attack on data systems
- Ransomware
- Theft

1.1.3 Lack of Clarity on Business Problem: First, one has to study the business challenge for which you want to apply data science solutions. Choosing a way to configure data setup equipment and perform data analysis before getting a clear picture of what business problem should be solved, seems ineffective.

1.1.4 Undefined KPIs and Metrics: Data scientists can design machine learning models and get accurate results with its help. However, there is a possibility that the metrics used do not serve the purpose of using the DS. The science of learning data not only involves the development of algorithms, but also requires a deeper understanding of other processes.

1.1.5 Difficulty in Finding Skilled Data Scientists: Lack of talent is another data science problem companies are facing. Businesses often struggle to find the right data team with in-depth knowledge and expertise in the field. As well as an in-depth understanding of ML and AI algorithms, professionals need to know more about the business vision of DS.

1.1.6 Getting Value Out of Data Science: Data experts believe that in order to support a business, the data analysis process needs to be faster and more consistent with the business during the decision-making process. Using DS allows you to build a culture of collaboration between team members and most importantly, gives your employees the ability to make better decisions.

II. HEADINGS

Data science is the integration of mathematics, computer technology and artificial intelligence (AI): These combined areas create new posts for companies like Google called data science. The data scientists team consists of mathematicians, computer scientists, AI scientists and experts in other relevant fields. Data-driven scientific discovery is an important emerging concept of computer use in areas that include social, service, Internet of Things, sensory networks, communications, biology, health care and cloud. Under this perspective, Data Science is the backbone of new research, from environmental to social. There are some related scientific challenges, ranging from data capture, creation, storage, search, sharing, modeling, analysis and visualization. Integration across all complex data-dependent resources for real-time decision-making, live-streaming of data, collaboration and maintaining the importance of collaboratively creating complex issues that need to be addressed. Data science covers the areas of Mathematics, Mathematics, Computer Science, Information Theory, Information Technology, machine learning and full functionality. It has become increasingly important to fully understand the big data sets and transform the data into a workable intelligence, be it business information, Government or the Web.

III. METHODOLOGY

3.1 Business Understanding: Before solving any problem in the Business domain it needs to be properly understood. Understanding business builds a tangible foundation, which leads to easy queries.

3.2 Analysis Comprehension: Based on the above business understanding one has to decide which analysis method to follow. The methods can be of 4 types: Descriptive method (current status and provided information), diagnostic method (statistical analysis, current and why), predictive method (predicts trends or opportunities for future events) and Prescriptive. method (how the problem should be solved).

3.3 Data Requirements: The analysis method selected above shows the required data content, formats and sources to be collected. During the data requirements process, one has to find answers to questions such as 'what', 'where', 'when', 'why', 'how' & 'who'.

3.4 Data Collection: Collected data can be obtained in any random format. Therefore, depending on the method selected and the result to be obtained, the data collected must be verified. Thus, if necessary one can collect additional data or discard non-essential data.

3.5 Data Understanding: Data comprehension answers the question "Does the data collected represent a problem to be solved?". Descriptive statistics calculate the metrics used in data to achieve content and content quality.This step may lead to a reversal of the previous step for correction.

3.6 Data Editing: Let's understand this by connecting this concept with two similes. One is a freshly washed bathvegetables and secondly only take the required items to eat on a plate during buffet.

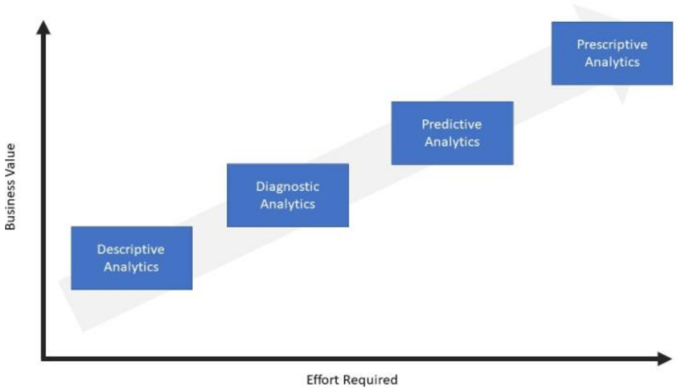
3.7 Modeling: Modeling determines whether the data prepared for processing is appropriate or needs additional completion and conservation.This section focuses on the construction of speculative / descriptive models.

3.8 Testing: Model testing is performed during model development. It checks the model level to be re-evaluatedand if it meets the needs of the business. It enters the diagnostic measurement phase (the model works as intended again where conversion is required) and the statistical assessment phase (ensures good data management and interpretation).

3.9 Shipping: As the model is successfully tested it is made ready for shipment in the business market.The feed section evaluates how well the model can withstand the external environment and is more efficient compared to others.

3.10 Answer: Feedback is a necessary goal that helps to refine the model and achieve its effectiveness and impact.The steps involved in responding define the review process, track the record, evaluate efficiency and refine refinement.

IV. FIGURES AND TABLES



4.1 Descriptive Analytics

The direct question answered by Descriptive Mathematics is, "What happened?" Looking back, understanding how the organization performed. This is especially helpful if the organization has launched a campaign and wants to see how it works. For example, a marketing campaign was launched in Jan 2019 and descriptive statistics can help track its effectiveness, and from there a decision can be made to determine whether the campaign needs to continue or be terminated.

4.2 Diagnostic Analytics

We do not just answer "What happened?". After knowing what has happened so far, we move on to the next step and that makes us understand the reason for what happened, "Why did it happen?". This is to allow us to learn from our actions so that success can be repeated and failure can be avoided.

Diagnostic Analytics is not an easy analysis to do. At this stage, there is a strong need for domain information. An analyst needs to understand a business very well, its process, regulations, policies, target market. The analyst is like a sleuth trying to find reasons.

4.3Predictive Analytics

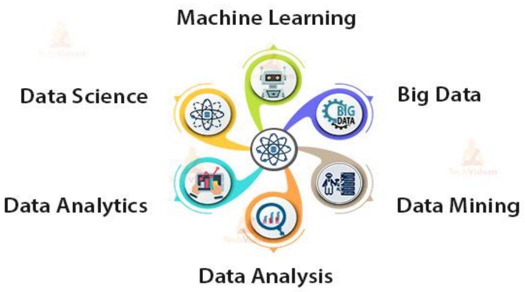
Predictive Analytics tries to answer the question, "What could happen?". Although from the name it sounds like we are making some really false predictions. In fact, in Predictive Analytics, we are looking at using machine learning algorithms to identify patterns. What kind of patterns?

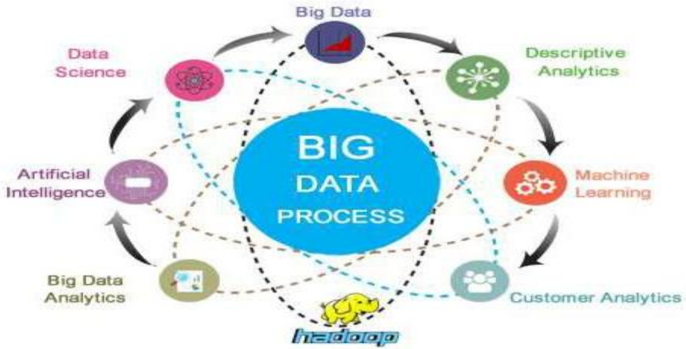
We find algorithms for working with data and find relationships between the target (something we wish to know in advance) and the features (something we will know before the goal becomes solid / final) that we are trying to apply. factors for determining whether a target can be.

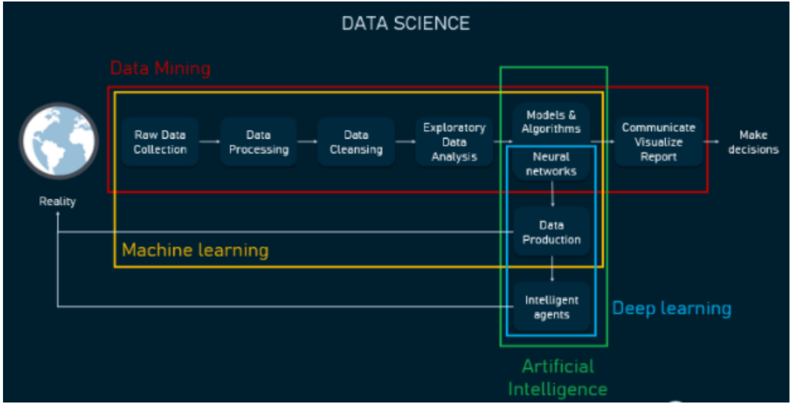
4.4 Prescriptive Analytics

Fixed statistics are much better than Predictive Analytics but how advanced are they? In Predictive Analytics, businesses can use to determine how often a customer is able to take a marketing offer, but Prescriptive Analytics, using statistical and mathematical methods to propose the next step, or answers the question, "What is the next best step?" It is possible to offer, a few options with possible consequences and businesses decide which is the best option you can make. Typically, a combination of the previous level of Analytics (Explanation, Diagnosis and Predictability) and possibly performance appraisal, game theory and much more.

4.5 FIELDS OF DATA SCIENCE







4.5.1 Big Data:

In today's world, a huge amount of data is building. Data is very important to organizations around the world. But the problem arises when it becomes difficult to manage and manage data using traditional techniques, in such cases, it is called big data.

4.5.2 Data Mining

Data mining is the process of discovering hidden patterns, trends, and information in data. In other words, we can say that the process of mining information i.e. useful information from a large amount of data i.e. green datasets.

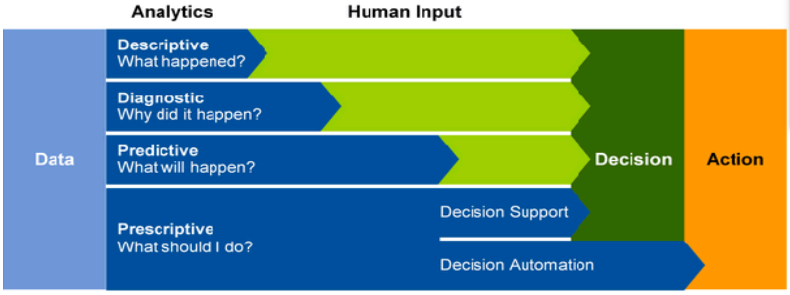
4.5.3 Data Analysis

In today's world, large amounts of data are available and can be collected from many sources such as surveys, interviews, surveys, etc. But this data is useless until we do not know how to turn this data into useful information. .Data Analysis is about how raw data is selected, analyzed and interpreted into logical and important conclusions that are easy to understand and use.

4.5.4 Data Analytics

Data Analytics can be defined as a method of data analysis. It can be considered a masterpiece of many different types of data analysis. Any type of raw data can be displayed in a variety of Data Matters that can help you find information that can improve things in many ways. Data analysis uses a variety of tools and techniques to collect data, analyze it and help to convert data in a way that is easy to understand and visualize.

4.5.5 Machine Learning



This machine follows the instructions given to people and does the right thing but what if one is able to train the machine to make its own decision. This is what we try to do using machine learning algorithms. Machine learning is used to build computer programs that can access data and learn from data itself.

V. Conclusion

As more and more fields emerge, the value of data science is also growing rapidly. Data science has influenced various areas. Its effects can be seen in many fields such as the retail industry, health care, and education. In the healthcare industry, new medicines and treatments are becoming increasingly available and there is a need for better patient care. With the help of data science techniques, the healthcare sector can find a solution that helps care for patients. Education is another field where the benefits of data science can be clearly seen. The latest technologies such as smartphones and laptops are now an integral part of the education system. With the help of data science, better opportunities are created for students who allow themselves to improve their knowledge. Data science is one of the growing fields. It has become an integral part of almost every sector. It provides the best solutions that help meet the challenges of ever-increasing need and a secure future. As the value of data science grows day by day, the need for data scientists grows. A data expert is the future of the world. Therefore, a data scientist should be able to provide good solutions that meet the challenges of all sectors. In order to do this, they need to have the right resources and programs to help them achieve their goal.

Acknowledgements

I would like to express my special thanks of gratitude to our Head of Department – Prof Mrs. Chandani Patel Ma'am, research coordinator Prof. Mrs. Neha Lodhe Ma'am and our research guide Prof. Mrs. krutika vartak ma'am who gave us the opportunity to this wonderful research on the topic "**Data Science**" , which also helped me in doing a lot of research and I came to know about so many new things. I am really thankful to them. Secondly, I would like to thank my research group's members in helping me carrying out my tasks at each phase of the development and contributing for the same equally and willingly, I also thank my parents and friends who helped me a lot in finalizing this research within limited time frame.

References

[1] Aggarwal, C.C. (ed.): Data Classification: Algorithms and Appli-cations. CRC Press, Boca Raton (2014)

[2] Allen, E., Allen, L., Arciniega, A., Greenwood, P.: Constructionof equivalent stochastic differential equation models. Stoch. Anal.Appl. 26, 274–297

[3] Anderson, C.: The End of Theory: The Data Deluge Makes theScientific Method Obsolete. Wired Magazine https://www.wired.com

[4] **Think Like a Data Scientist**, B. Godsey, **Manning** (2017)

[5] **Doing Data Science**, C. O'Neill & R. Schutt, **O'Reilly** (2013)
[6] Cao, L. B. & Yu, P. S. (2009) Behavior Informatics: An Informatics Perspective for Behavior Studies. IEEE Intel-
[8] Iwata, S. C. (2008) Editor's Note: Scientific 'Agenda' of Data Science. Data Science Journal 7, pp 54–56.
[9] Liu, L., Zhang, H., Li, J. H., et al. (2009) Building a Community of Data Scientists: an Explorative Analysis. Data Science Journal 8, p 24
[10] EMC (2011) Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field. Retrieved from the World Wide Web November 11, 2014: http://www.emc.com/collateral/about/news/emc-data-science-study-wp.pdf
[11] Donoho, D.: 50 Years of Data Science. http://courses.csail.mit.edu/18.337/2015/docs/ 50YearsDataScience.pdf (2015)
[12] Dyk, D.V., Fuentes, M., Jordan, M.I., Newton, M., Ray, B.K.,Lang, D.T., Wickham, H.: ASA Statement on the Role of Statisticsin Data Science. http://magazine.amstat.org/blog/2015/10/01/ asa-statement-on- the-role- of-statistics-in-data-science/ (2015)